PHYLOGENETIC SIGNAL, PHASE TRANSLATIONS AND LIMITS TO RESOLVING DEEP DIVERGENCES

E Mossel and M Steel

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2004/2 JANUARY 2004

Phylogenetic signal, phase transitions and limits to resolving deep divergences

Elchanan Mossel*
Statistics
U.C. Berkeley
mossel@stat.berkeley.edu

Mike Steel †
Biomathematics Research Centre
University of Canterbury
m.steel@math.canterbury.ac.nz

January 26, 2004

Abstract

In this chapter we review some recent results that shed light on a fundamental question in molecular systematics: how much phylogenetic 'signal' can we expect from characters that have evolved under some Markov process? We describe some explicit and easily-computable bounds on this information, both for finite-state and infinite-state models.

1 Introduction

As biologists delve deeper into the evolutionary history of life they often find that sequence data provides conflicting or unclear phylogenetic information. For DNA sequences that have a high site substitution rate the problem of *site saturation* is well known, whereby certain sequences are essentially random with respect to each other due to the number of substitutions that have occurred during their evolution from a common ancestral sequence. For other sorts of data - such a gene order data, where genomes have undergone much reshuffling - a similar eventual randomization and loss of information also occurs.

The phenomenon of randomization, and the rate at which it occurs, have been well studied in the probability literature - see for example Diaconis [12]. In this setting it is often useful to regard the stochastic process as a random walk on a group. For example, card shuffling, or gene order rearrangement may be viewed as a random walk on the symmetric group on n elements, while site substitution in DNA sequences of length k may be regarded as a random walk on the group $(\mathbb{Z}_2 \times \mathbb{Z}_2)^k$ (this was first noted by Evans and Speed [15]). An alternative setting is to consider a random walk on a finite regular connected graph, and most of the examples we have just mentioned can also be viewed from this perspective. Either setting - a random walk on a group, or a random walk on a graph - is just a special type of ergodic Markov chain, for which the usual questions

*Supported by a Miller fellowship in Computer Science and Statistics, U.C. Berkeley.
†Corresponding author. Thanks to the New Zealand Institute for Mathematics and its Applications (Phylogenetic Genomics Programme and Maclauren Fellowship)

arise, such as what is the limiting distribution, and how fast does the chain approach this limit? Often there is an abrupt transition from non-random to random in a sense that can be formalized and proved. For example, with binary sequences of length k (where k is large) under a model of independent site substitution, this transition occurs when each site has undergone approximately $\frac{1}{4}\log_e(n)$ substitutions - beyond this point the derived sequence quickly becomes essentially random with respect to the first (for a precise rendition of this statement see [12], Theorem 3, p. 28).

While these questions have been well understood for Markov chains, they have been less thoroughly investigated for the more general setting of Markov processes on trees.

The situation here is interesting for the following reason - as the tree gets larger each leaf tends to become further from the root (and so conveys less information about the ancestral root state) yet the number of leaves also gets larger. It is, a priori, not clear whether the gain in information provided by more leaves compensates for the losses experienced by each leaf. This question is also familiar in biology - does the sampling of more species provide a strategy for coping with site saturation? As we will see, these questions are relevant not just for reconstructing ancestral character states, but also for inferring phylogenetic trees.

Evolution processes may be often viewed as Markov processes on trees. These processes are in turn a special family of Markov random fields on trees, the study of which is an important branch of *statistical physics* - see [21] for general background and [25, 14, 38, 42, 32] for results regarding Markov processes on trees. The theory of Markov random fields (and processes) on trees is used to investigate problems such as ancestral reconstruction of states, which is familiar in both biology and physics. In contrast, the problem of reconstructing the tree topology, which is well-studied in biology, seems not to have been addressed in the statistical physics literature.

In this chapter we survey some of the recent advances in the information-theoretic treatment of Markov processes on trees. We begin by dealing with Markov processes on a fixed (small) state space - for example nucleotide sequence data. Here we describe information-theoretic limits that place bounds on the extent to which ancestral states and deep divergences can be resolved from sequence data. We also consider the question of how much sequence information is required to accurately reconstruct a tree, a question where there remains an interesting unresolved issue. We then turn to the analysis of characters on state spaces that are large or infinite, and which exhibit a somewhat different (and more tractable) behavior. Along the way we will indicate how such character data may be relevant to the analysis of genomic data such as gene order.

2 Preliminaries

In this section we describe some background and notation concerning phylogenetic trees and Markov processes on trees - readers familiar with these topics may wish to skim over this material.

2.1 Phylogenetic trees

Throughout this chapter X is a finite set and we will let $n = |X|$. A *phylogenetic X–tree* (or more, briefly, a *phylogenetic tree*) is a tree $T = (V, E)$ having leaf set X, and for which the interior vertices are unlabelled and of degree at least 3. If in addition each interior vertex has degree exactly 3 we

say that T is *trivalent.*

Two phylogenetic X–trees T and T' are regarded as equivalent if the identity map on X, regarded as a bijection from the set of leaves of T to the leaves of T' extends to a graph isomorphism between the two trees. Thus, for example, there are precisely three trivalent (and one non-trivalent) phylogenetic X–trees for any set X of size 4.

We are also interested in *rooted* phylogenetic X–trees. Briefly, a rooted phylogenetic tree is obtained from a phylogenetic tree by either distinguishing some interior vertex as a root, or by subdividing an interior edge and calling the new degree-two vertex a root. We denote the root of a rooted phylogenetic tree T by ρ, and direct all edges away from the root. For a rooted phylogenetic tree T its *topology* is the associated unrooted phylogenetic tree (obtained by suppressing the root, and if it is of degree 2 identifying its two incident edges). A rooted phylogenetic tree is said to be *binary* if each non-leaf vertex has precisely two outgoing arcs. Thus a phylogenetic tree is binary precisely if its topology is trivalent. For more background on the mathematics of phylogenetic trees the reader is referred to [51].

2.2 Markov processes on trees

Let $\mathcal{C}$ be the set of character states (such that $\mathcal{C} = \{A, C, G, T\}$ or $\mathcal{C} = \{20 \text{ amino acids}\}$). In keeping with biological convention we will often refer to a an site aligned across a set of species X as a *character* on X; mathematically it is simply a function from X to $\mathcal{C}$. To model the evolution of characters on a rooted phylogenetic tree T by a Markov process we associate to each directed edge e of T a matrix $M(e)$ of transition probabilities, and to the root vertex of T we associate a distribution π of states (see [13] or [53] for a more formal description of the model).

Many of the standard models in biology satisfy $M(e) = \exp(t(e)Q)$, where $Q = (q_{i,j})_{i\in\mathcal{C}, j\in\mathcal{C}}$ is the transition rate matrix and $t(e)$ represents the 'length' of the edge e over which the Markov process operates. Furthermore, π is generally taken to be the equilibrium distribution that satisfies $\pi Q = 0$, so as to induce a stationary Markov process.

The simplest 2–state model is the symmetric *Cavender-Farris-Neyman (CFN) model*

$$Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},$$

For this model the probability $p(e)$ of a substitution on any edge e of the tree is given by

$$p(e) = \frac{1}{2}(1 - \exp(-2t(e)). \tag{1}$$

With 4 states a slightly more general class of models is the Tajima and Nei's 'equal input' model

$$Q = \begin{pmatrix} -(a+b+c) & a & b & c \\ d & -(b+c+d) & b & c \\ d & a & -(a+c+d) & c \\ d & a & b & -(a+b+d) \end{pmatrix}.$$

In case $a = b = c = d (= r$, say) this is known as the *Jukes-Cantor model*:

$$Q = \begin{pmatrix} -3r & r & r & r \\ r & -3r & r & r \\ r & r & -3r & r \\ r & r & r & -3r \end{pmatrix}. \tag{2}$$

Both of these models lead to reversible Markov processes. See [18] for various other families of substitution matrices Q appearing in biology.

A further embellishment of most contemporary models of nucleotide substitution is the inclusion of *site specific rates*. That is, one has a distribution $\mathcal{D}$ on some real-valued parameter (the 'rate' of evolution of a site) and each site i in the sequence evolves at a rate λ_i that is chosen independently from this distribution. We refer to the distribution that assigns rate 1 to each site with probability 1 as the *degenerate distribution*.

The substitution process is therefore defined by a transition rate matrix Q, a distribution $\mathcal{D}$ of site specific rates, a rooted phylogenetic tree $T = (V, E, \rho)$, a collection of edge lengths $t : E \to \mathcal{R}_+$ and a probability distribution π on the states at the root vertex of T.

A *configuration* $\sigma : V \to \mathcal{C}$ is a labeling of the vertices of T by $\mathcal{C}$. We will write σ_v for the value of σ at the vertex $v \in V$. The distribution of σ_ρ is given by π. If u is v's parent, then the conditional distribution of σ_v given σ_u at site i is given by the matrix $M(e) = \exp(\lambda_i t(e)Q)$, where $e = (u, v)$. We will denote the collection of leaves of the tree T by ∂T and the value of a configuration σ at the leaves by σ_∂ (which is a character on X - that is, a function from X into the set $\mathcal{C}$).

3 Information-theoretic bounds: ancestral states and deep divergences

In this section we describe explicit and easily computable upper bounds on the information that extant sequences provide concerning (i) ancestral sequences and (ii) the branching pattern deep inside a tree. These bounds are in a sense the simplest bounds that can be put on the reconstruction of ancestral data

For a leaf v, let path(v) be the set of edges on the path connecting v to the root ρ, and let

$$t(v) = \sum_{e \in \text{path}(v)} t(e).$$

The *molecular clock assumption* is that $t(v)$ takes the same value for each v; we do not make this assumption anywhere in this chapter, even though we will refer to sums of $t(e)$ values as (elapsed) 'time'.

Let π be the prior distribution of the root character, and let

$$\Delta = \sup_f \mathbb{P}[f(\sigma_\partial) = \sigma_\rho], \tag{3}$$

be the optimal probability of reconstructing the value of σ_ρ given σ_∂, where the sup is taken over all functions. Assuming that the parameters of the model (i.e. T, the $t(e)$ values and the root state

distribution π) are known, it follows from a classic result (see for example Theorem 17.2 of [24]) that an optimal choice of f is the maximum posterior probability (MAP) estimator - that is, given σ_∂ one select the root state(s) j to maximize

$$\mathbb{P}[\sigma_\partial | \sigma_\rho = j] \cdot \pi[\sigma_\rho = j]$$

- a task that can be carried out by an efficient (polynomial-time in n) dynamic programming approach.

It also follows from standard information-theoretic theory (Theorem 17.3 of [24]) that the following lower bound on Δ applies:

$$\Delta \geq 2^{-H(\sigma_\rho | \sigma_\partial)} \tag{4}$$

where $H(\sigma_\rho | \sigma_\partial)$ is the *conditional entropy* of σ_ρ given σ_∂ is defined by

$$H(\sigma_\rho | \sigma_\partial) = -\sum_{i,\sigma} \mathbb{P}[\sigma_\rho = i, \sigma_\partial = \sigma] \log_2(\mathbb{P}[\sigma_\rho = i | \sigma_\partial = \sigma]).$$

However our main interest here is in providing explicit upper bounds on Δ, which we now describe. As the rate of substitution increases and/or the temporal separation of the root of the tree from the leaves increases, we would expect it to become increasingly difficult to recover the root state – a phenomenon well known to biologists as 'site saturation'. However it will be important (particularly for later results) to quantify this rate of decay of information. The following result, which is a slight extension of a result from [40], follows by easy adaptations of coupling arguments appearing earlier in statistical physics, see, e.g., [38]. We let $M_{\mathcal{D}}(x) = \mathbb{E}[e^{\lambda x}]$ the moment generating function of the site specific rate distribution $\mathcal{D}$. Note that, for the degenerate site specific rate distribution we have $M_{\mathcal{D}}(x) = e^x$.

Theorem 3.1. *Consider a Markov model on a tree T, with transition rate matrix Q, edge lengths $t(e)$ (for each edge e of T), and site specific rate distribution $\mathcal{D}$. Let*

$$q_j = \min_{i \neq j} q_{i,j}, \quad q = \textstyle\sum_j q_j. \tag{5}$$

Then the optimal reconstruction probability Δ for the root state satisfies

$$\Delta \leq \max_i \pi[\sigma_\rho = i] + \sum_{v \in \partial T} M_{\mathcal{D}}(-qt(v)). \tag{6}$$

Note that the first term in (6) is precisely the estimate one would make if one had no knowledge of the character states at the leaves of T. Thus Theorem 3.1 says that the improvement over this 'trivial' method decays as the expected exponential of $-qt(v)$. Notice also that Theorem 3.1 assumes that T and the values $t(e)$ are all known exactly - if they are not, then the bound on Δ described applies *a fortiori*.

The proof of Theorem 3.1 utilizes the method of coupling (see e.g. [1] for background on coupling for Markov chains) and arguments from the theory of percolation (see [38], and [47, 3] for background). We outline this argument now. First, we establish the result for the special case of constant site specific rate, where each site is assigned rate λ with probability 1. The substitution rate from state i to state j is given by $q_{i,j}$. Recalling (5), we may define the process equivalently as follows. Given the current state i,

(**J1**) jump to state j with rate λq_j;

(**J2**) jump to state j with rate $\lambda(q_{i,j} - q_j)$.

The crucial point here is that (**J1**) is performed independently of the state i. For edge $e = (u, v)$, let $D(e)$ be the event that a transition of type (**J1**) occurs along the edge e. Note that the events $D(e)$ are independent for different edges and that $\mathbb{P}[D(e)^c] = \exp(-q\lambda t(e))$. Moreover, conditioned on $D(e)$, σ_v is independent of σ_ρ. For a leaf v, let $D(v)$ be the event that transition of type (**J1**) occurs along an edge $e \in \text{path}(v)$. Then

$$\mathbb{P}[D(v)^c] = \prod_{e \in \text{path}(v)} \mathbb{P}[D(e)^c] = \prod_{e \in \text{path}(v)} e^{-q\lambda t(e)} = e^{-q\lambda t(v)}.$$

Finally, let D be the event that $D(v)$ holds for all leaves $v \in \partial T$. Then

$$\mathbb{P}[D^c] \leq \sum_{v \in \partial T} \mathbb{P}[D(v)^c] = \sum_{v \in \partial T} e^{-q\lambda t(v)}. \tag{7}$$

Note that conditioned on D, σ_∂ and σ_ρ are independent.

To prove the bound on reconstruction (6), note that if we are not given σ_∂ (or any other information on σ_ρ), then the best reconstruction function f satisfies $f \equiv j$, where j maximized $\pi[\sigma_\rho = i]$ over all i, and this function has success probability $\max_i \pi[\sigma_\rho = i]$. Now let f be any reconstruction procedure and note that, condition on the event D, σ_ρ is independent of σ_∂ and therefore

$$\begin{aligned} \mathbb{P}[f(\sigma_\partial) = \sigma_\rho] &\leq \mathbb{P}[D^c] + \mathbb{P}[D]\mathbb{P}[f(\sigma_\partial) = \sigma_\rho | D] \\ &\leq \mathbb{P}[D^c] + \mathbb{P}[D] \max_i \pi[\sigma_\rho = i] \leq \mathbb{P}[D^c] + \max_i \pi[\sigma_\rho = i], \end{aligned}$$

and so

$$\mathbb{P}[f(\sigma_\partial) = \sigma_\rho] \leq \max_i \pi[\sigma_\rho = i] + \sum_{v \in \partial T} e^{-q\lambda t(v)}. \tag{8}$$

Now, consider the case of a general site specific rate distribution $\mathcal{D}$. Clearly, Δ is the expected value (with respect to $\mathcal{D}$) of the conditional probability $\mathbb{P}[f(\sigma_\partial) = \sigma_\rho | \lambda]$ which we may identify with the left-hand side of (8). Consequently,

$$\Delta \leq \mathbb{E}_{\mathcal{D}}[\max_i \pi[\sigma_\rho = i]] + \mathbb{E}_{\mathcal{D}}[\sum_{v \in \partial T} e^{-q\lambda t(v)}] = \max_i \pi[\sigma_\rho = i] + \sum_{v \in \partial T} M(-q\lambda t(v))$$

as required. ■

Example. To illustrate Theorem 3.1 let us consider the simplest model on four states, namely the Jukes-Cantor model (2) with a degenerate site specific rate distribution and a molecular clock. For this model the equilibrium distribution for states is uniform, so it is natural to take $\pi[\sigma_\rho = i] = \frac{1}{4}$ for all four choices of i. Now suppose we wish to infer the ancestral state at a vertex in a tree that was present t years ago, using the states observed now amongst the n extant descendant species. Theorem 3.1 provides the following bound on Δ:

$$\Delta \leq \frac{1}{4} + ne^{-qt},$$

and we may identify the product $\frac{3}{4}qt$ with the expected number of substitutions that occur on any path from the root to a leaf. For example, if the substitution rate is constant at (say) 1 substitution per million years, and we have a tree with $n = 100$ leaves whose root is at least 10 million years in the past then $\Delta \leq \frac{1}{4} + 0.0002$ so a character tells us virtually nothing to help us estimate the state that occurred at the root. □

Notice that some restriction must be placed on the entries of Q for a bound such as that given by (6) to apply. For example, consider a process with three states, with $\pi[\sigma_\rho = i] = 1/3$ for each value of i, and with

$$Q = \begin{pmatrix} -2r & r & r \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

for which $q = 0$. Then it is easily checked that $\Delta \geq 1/2$, and yet $\max_i \pi[\sigma_\rho = i] = 1/3$.

However Theorem 3.1 can be extended to provide some (exponential-decay) bounds similar to (6) for certain choices of Q for which $q = 0$. A case in point is the class of 'covarion-type' models (see [20], [46],[59]) in which each state can either be in an 'on' mode or an 'off' mode. A state that is 'on' is free to change to other 'on' states, or to turn 'off' (at various rates), while a state that is 'off' is only free to turn 'on' (at some rate). For two base states and therefore a total of 4 states, namely $0_{\text{on}}, 1_{\text{on}}, 0_{\text{off}}, 1_{\text{off}}$ the corresponding rate matrix Q can be written as:

$$Q = \begin{pmatrix} -(r_1 + u) & r_1 & u & 0 \\ r_2 & -(r_2 + u) & 0 & u \\ v & 0 & -v & 0 \\ 0 & v & 0 & -v \end{pmatrix}. \tag{9}$$

and for this matrix it is immediately clear that $q = 0$.

In order to obtain bounds for such models, it is better to apply the coupling argument directly to the matrices $M(e)$. Given any real matrix A let $m_j(A) = \min_i A_{i,j}$ and $m(A) = \sum_j m_j(A)$. Write $m_j(e)$ for $m_j(M(e))$ and $m(e)$ for $m(M(e))$. On the edge e, the transition process can be described equivalently as follows: Given the current state i,

(J1) jump to state j with probability m_j;

(J2) jump to state j with probability $m_{i,j} - m_j$.

Note that, as before, **(J1)** is performed independently of the state i. Repeating the above argument we thus obtain the following bound on the reconstruction probability

$$\Delta \leq \max_i \pi[\sigma_\rho = i] + \sum_{v \in \partial T} \prod_{e \in \text{path}(v)} (1 - m(e)). \tag{10}$$

For a given tree and substitution matrices we may apply bound (10) directly. However, unlike Theorem 3.1, here it is not enough to know for all leaves the total time elapsing from the root. Instead, all the edge lengths are needed.

More can be said if the process described by Q is ergodic (maybe with 0 entries) and it is assumed that the length of all branches is at least ϵ. Let $\alpha = \sqrt{1 - m(\exp(\epsilon Q))}$ and note that $\alpha < 1$.

Note by the coupling argument above that if A and B are two stochastic matrices, then $1 - m(AB) \leq (1 - m(A))(1 - m(B))$. Thus, if $t > \epsilon$, then

$$1 - m(\exp(tQ)) \leq 1 - m\left(\left(\exp(\epsilon \lfloor \frac{t}{\epsilon} \rfloor Q)\right)\right) \leq (\alpha^2)^{\lfloor \frac{t}{\epsilon} \rfloor} \leq (\alpha^2)^{\frac{t}{2\epsilon}} = \alpha^{t/\epsilon}.$$

Substituting this into (10) we obtain that

$$\Delta \leq \max_i \pi[\sigma_\rho = i] + \sum_{v \in \partial T} \alpha^{t(v)/\epsilon}.$$

Note the similarity between this expression and the one in Theorem 3.1. In particular, in order to apply this bound it suffices to know for each leaf the total time elapsed from the root.

3.1 Reconstructing deep divergences

Theorem 3.1 allows one to place bounds on the extent to which sequences can resolve a divergence event deep inside a phylogeny. Consider for example four monophyletic groups of taxa for which we have aligned sequences of length k. We may wish to determine which of the three possible phylogenetic trees connect these four groups, as illustrated on the left of Fig. 1.

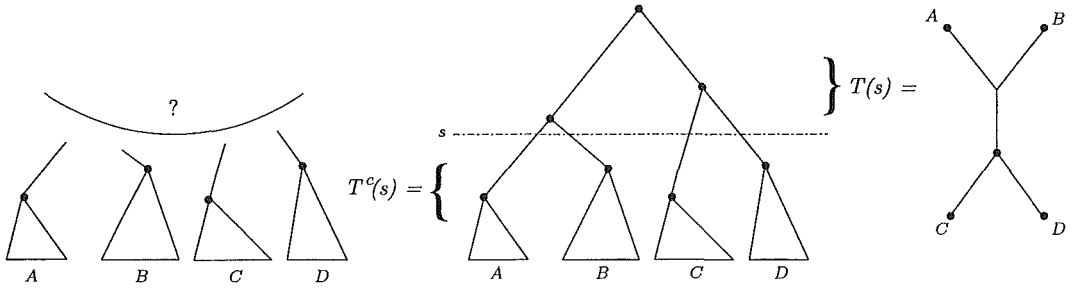

Figure 1: *Left:* An example of a deep divergence involving four subtrees. *Centre and Right:* The tree $T(s)$ and forest $T^c(s)$

Clearly, it will only help us in this task if we know the tree topologies of each of the four monophyletic groups together with their $t(e)$ values. Each sequence site provides a quantum of information concerning the 'deep' tree structure (i.e. which of the three possible phylogenetic trees connect the four subtrees) and it is possible to explicitly bound the information that the entire sequences provide concerning this divergence. In this way one can set explicit lower bounds on the number of sites would be needed in order to resolve a deep divergence. One such bound was described, for the CFN model, in [52]. Here we describe a more general approach from [40] that applies to a wider range of models and settings.

Let $T(s)$ denote the topology of tree T up to time s from the root, and let $T^c(s)$ denote the forest consisting of the subtrees from time s to the present (including the associated edge lengths). In other words, $T(s)$ describes all divergences up to time s, while $T^c(s)$ describes all divergences (and their relative separations) from time s, as illustrated on the center and right of Fig. 1.

Consider the problem of reconstructing $T(s)$ (given $T^c(s)$) from a sequence of characters that are generated by a common Markov process on T, where the prior distribution on $T(s)$ is given by a measure μ. The prior μ is on $T(s)$ with its edge lengths. However, for a tree topology T, we will write $\mu[T(s) = T]$ for the prior probability that the topology of $T(s)$ is given by T.

Note that in the following result (Theorem 3.2) we do not need to assume independence between sites that evolve according to this process on T.

Let us denote a sequence of k identically generated configurations by $\sigma^1, \ldots, \sigma^k$. We will also denote the values of the configuration σ^i at the leaves by σ^i_∂. Similarly, we denote by σ^i_ρ the value of the configuration σ^i at the root ρ. Suppose furthermore, that the characters evolve as in Theorem 3.1 with substitution matrix Q; and we have a site specific rate distribution $\mathcal{D}$. Let $\Delta^T(s)$ be the probability of reconstructing, given $T^c(s)$ (with its associated $t(e)$ values) the tree topology up to time s,

$$\Delta^T(s) = \sup_f \mathbb{P}[f((\sigma^j_\partial)_{j=1}^k) = T(s)|T^c(s)]. \tag{11}$$

The sup is taken over all functions, and as before, the optimal choice of f is the maximum posterior probability (MAP) estimator, which given $(\sigma^j_\partial)_{j=1}^k$ selects a tree T' to maximize

$$\int 1_{\{T(s)=T'\}} \mathbb{P}[(\sigma^j_\partial)_{j=1}^k | T(s)] d\mu(T(s)).$$

Clearly the probability of reconstructing T from $(\sigma^j_\partial)_{j=1}^k$ is less or equal to $\Delta^T(s)$; this latter quantity, which is the probability of correctly determining the 'deep' part of the tree, can be bounded as follows.

Theorem 3.2. *Suppose that k sites evolve under a Markov process with a site specific rate distribution $\mathcal{D}$. Then, for any $s > 0$ we have:*

$$\Delta^T(s) \leq \max_T \mu[T(s) = T] + k \sum_{v \in \partial T} M_{\mathcal{D}}(-q(t(v) - s)), \tag{12}$$

where q is given by (5).

Outline of the proof. The argument follows similar lines to the proof of Theorem 3.1. For character i we say that event D_i occurs if, for all $v \in \partial T$ there exists a time $t \geq s$ at which a transition of type (**J1**) occurs at least once on the path connecting v to the root of the component of $T^c(s)$ that contains v. By the proof of Theorem 3.1 it follows that

$$\mathbb{P}[D_i^c | \lambda_i] \leq \sum_{v \in \partial T} e^{-\lambda_i q(t(v) - s)},$$

where λ_i is the rate (chosen from $\mathcal{D}$) that site i evolves at. Consequently,

$$\mathbb{P}[D_i^c] \leq \sum_{v \in \partial T} M_{\mathcal{D}}(-q(t(v) - s)),$$

and so, by the Bonferroni inequality,

$$\mathbb{P}[(\cap_{i=1}^k D_i)^c] \leq k \sum_{v \in \partial T} M_{\mathcal{D}}(-q(t(v) - s)).$$

Now, conditional on $\cap_{i=1}^k D_i$, the two random variables $(\sigma_s^i)_{i=1}^k$ and $(\sigma_\partial^i)_{i=1}^k$ are independent, and therefore, $T(s)$ and $(\sigma_\partial^i)_{i=1}^k$ are independent. As in Theorem 3.1 we conclude that

$$\Delta^T(s) \leq \mathbb{P}[(\cap_{i=1}^k D_i)^c] + \mathbb{P}[(\cap_{i=1}^k D_i)] \max_T \mu[T(s) = T] \leq k \sum_{v \in \partial T} M_{\mathcal{D}}(-q(t(v)-s)) + \max_T \mu[T(s) = T],$$

as required. □

Example. To illustrate Theorem 3.2 let us consider again the Jukes-Cantor model (2), with a degenerate site specific rate distribution and molecular clock. Suppose we have four monophyletic groups of taxa - each with 100 extant species, and with a well-specified tree with edge lengths - and we wish to determine which of the three possible trees (choices for $T(s)$) describes how the trees are joined ancestrally (as in Fig. 1). In the absence of any prior information it is natural to take $\mu[T(s) = T'] = \frac{1}{3}$ for each of the three possible trivalent trees T'. Suppose it is believed that all four lineages existed as far back as (at least) 1 billion years ago, and taking (for example) a site substitution rate $(3r)$ of one substitution per fifty million years, we have for any leaf v that $qt(v) = 4rt(v) = \frac{4}{3} \cdot (3r)t(v) = \frac{4}{3} \cdot 20$. Theorem 3.2 then gives $\Delta^T(s) \leq \frac{1}{3} + 100ke^{-26.7}$ which implies that at least 700 million sites (!) would be required in order to have any hope of estimating the ancestral divergence with probability more than about 0.5. □

Remarks

(1) As noted above, Theorem 3.2 applies even when the sequence sites are not independent. It is possible to extend this theorem further to allow the sites to evolve according to different Markov processes.

(2) In order to get a feeling for the asymptotic behavior of (12), fix s and assume that the tree has $n = e^{\beta t}$ leaves, all at time t. Here we take the asymptotics where $t \to \infty$ (and therefore $n \to \infty$), while s, q and β are all constants. Also we assume a degenerate site specific rate distribution. Then

$$\sum_{v \in \partial T} e^{-q(t(v)-s)} = \exp(sq)\exp(-t(q-\beta)).$$

Therefore if $q > \beta$, then by (12) if we want to reconstruct the topology up to time s with high probability, i.e., $\Delta^T(s) \geq \max_T \mu[T(s) = T] + n^{-o(1)}$, then we need that

$$k \geq \exp(t(q - \beta - o(1))) = n^{q/\beta - 1 - o(1)}.$$

So the number of characters needed is polynomial in n.

□

3.2 Connection with information theory

Similar bounds to the ones we have described so far can also be stated and derived using classical information theory. First we briefly recall the concept of mutual information. For a random variables X, and Y the *mutual information* between X and Y is defined by

$$I(X,Y)(= I(Y,X)) := \sum_{x,y} \mathbb{P}[X = x, Y = y] \log_2 \left(\frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[X = x]\mathbb{P}[Y = y]} \right).$$

Formally, $I(X,Y)$ is the Kullback-Leibler separation of the joint distribution of X,Y and the product distribution of X and Y. Consequently, $I(X,Y) \geq 0$ with equality if and only if X and Y are independent. Informally $I(X,Y)$ measures the amount of information that Y carries about X (or conversely that X carries about Y). When $I(X,Y)$ is small then the best method for inferring Y from X does little better than the best method that simply ignores X - a precise formalization of this claim is Fano's inequality (see [10] for more details).

The quantity I has some generic properties that make it useful for analyzing the information loss of Markov processes. For example, suppose that X,Y and Z be random variables such that X and Z are independent given Y. Then $I(X,Z) \leq \min\{I(X,Y), I(Y,Z)\}$ (the 'data processing lemma') and $I((X,Z),Y) \leq I(X,Y) + I(Z,Y)$ (the 'subadditivity property'). By exploiting these properties one can derive information-theoretic analogues of Theorems 3.1 and 3.2 which we will now briefly describe. For convenience we will deal just with the degenerate site distribution in both cases. In the setting of Theorem 3.1 it can be shown that

$$I(\sigma_\partial, \sigma_\delta) \leq \log_2 |\mathcal{C}| \sum_{v \in \partial T} e^{-qt(v)}.$$

Similarly, in the setting of Theorem 3.2 it can be shown that

$$I(T(s); (\sigma_\partial^j)_{j=1}^k | T^c(s)) \leq k \sum_{v \in \partial T} e^{-q(t(v)-s)}.$$

For further details, and applications of these results, see [40].

4 Phase transitions in ancestral state and tree reconstruction

There is an interesting transition in the behavior of Markov models of character evolution on trees. This has been well studied in statistical physics and in information theory, in the context of broadcasting on trees. But it is also relevant to biology - particularly in attempting to recover information (ancestral states, branching order) deep within a tree, from observing the character states at the leaves.

The transition is most easily explained, and has been most studied for the case of the 2–state symmetric process (the CFN model described above).

To illustrate this transition between what is called the 'ordered' and 'unordered' phases of a Markov process on a tree, suppose we have a rooted binary phylogenetic tree T that has $n = 2^m$ leaves that are at distance m from the root vertex, as indicated in Fig. 2.

Under the CFN model (and with a degenerate site specific rate distribution) let

$$\theta(e) := \det(M(e)) = \det(t(e)Q) = \exp(t(e)\mathrm{tr}(Q)) = e^{-2t(e)},$$

where the penultimate inequality is Jacobi's identity (where tr denotes matrix trace). By (1) we have $\theta(e) = 1 - 2p(e)$. Now suppose that each edge of T has the same $t(e)$ value, say t, and thereby the same $\theta(e)$ value, namely $\theta = \exp(-2t)$.

Let us further suppose that the distribution π of states at the root is uniform (i.e. a fair coin toss) and that we wish to use the states $\sigma_\partial = (\sigma_\partial^i)$ at the leaves of T to estimate the state σ_ρ at

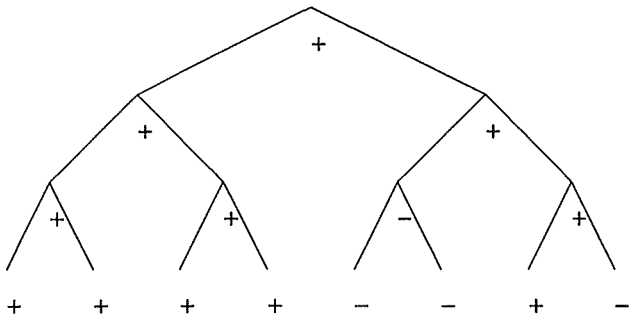

Figure 2: A character of the CFN process on a binary phylogenetic tree on $8 = 2^3$ leaves at distance 3 from the root

the root. This gives rise to an interesting contest as m (the height of the tree) increases - firstly, each leaf is becoming increasingly far from the root, and so the information that it carries about the ancestral root state decays to 0 with increasing m. On the other hand, the number of leaves is grows (exponentially) with m, and so although each leaf carries less information, it might be hoped that together they compensate for their individual losses. Which factor wins out depends critically on the value of θ. Evans *et al.* [14] established that, for $2\theta^2 < 1$ the mutual information $I(\sigma_\theta, \sigma_\rho)$ converges to 0, as m tends to infinity (this result was first proven independently by [5] in a different formulation). Thus, eventually (as the root becomes increasingly 'deep' in the tree) it becomes impossible to estimate the root state with any better success than a blind guess, when θ lies in this region. On the other hand, when $2\theta^2 > 1$ then $I(\sigma_\theta, \sigma_\rho)$ is bounded away from 0, so that information about the root 'survives' to the leaves, no matter how large the tree grows. In this case maximum likelihood estimation or majority rule estimation (i.e. select the root state that corresponds to the majority state at the leaves) suffices to recover some information right up to (but not including [44]) the critical value $2\theta^2 = 1$.

Notice that this critical value translates to a common $t(e)$ value of $t = \frac{1}{4}\log(2)$ and thereby to a common $p(e)$ value of $p = \frac{1}{2}(1 - \frac{1}{\sqrt{2}})$.

Curiously, the maximum parsimony approach for ancestral state reconstruction (i.e. select the root state that requires the fewest transitions to account for the leaf states) recovers information under the CFN model for values of p only up to $\frac{1}{8}$ [7].

The situation for r-states models and for non-symmetric 2-state processes is more subtle. There is not any general criteria for deciding when the mutual information $I(\sigma_\theta, \sigma_\rho)$ is converging to 0 and when is it bounded away from 0. In fact, such criteria do not even exist for symmetric processes on more than 2 states or for general processes on 2 states. In the general setting, there are various conditions which imply that mutual information either converges to 0 or is bounded away from 0, however these conditions are not sharp. We describe an example of both types of conditions now.

Suppose that $M(e) = M$ for all e. Since M is a stochastic matrix, 1 is an eigenvalue of M. Let $\{1 = \lambda_1, \ldots, \lambda_r\}$ denote the set of eigenvalues of M and let $\theta = \max\{|\lambda_2|, \ldots, |\lambda_r|\}$ (note that for the CFN model, this is consistent with the previous definition of θ). It is known ([28, 42]) that for any M if $2\theta^2 > 1$ then $I(\sigma_\theta, \sigma_\rho)$ is bounded away from zero. This result is not tight in general (see

[36, 42, 27]).

In order to illustrate the last bound consider the Jukes-Cantor model (2). Note that the eigenvalues of Q are 0 (with multiplicity 1) and $-3r$ (with multiplicity 3). Thus if $M = \exp(tQ)$, then the eigenvalues of M are 1 and e^{-3rt}. Therefore, if the stochastic matrix $M = \exp(tQ)$ satisfies

$$2e^{-6rt} > 1 \tag{13}$$

then $I(\sigma_\partial, \sigma_\rho)$ is bounded away from zero.

In the other direction, various conditions are derived in [36, 42, 32, 31] that imply that $I(\sigma_\rho, \sigma_\partial)$ converges to 0 for various processes. The simplest of these conditions is given in [36] - this condition is closely related to the one given in the previous section. The results in [42, 32, 31] give sharper bounds for symmetric processes on more than 2 states and for general 2–state processes.

4.1 The logarithmic conjecture

Suppose we generate k characters independently and according to the CFN model (with degenerate site specific rate distribution), and ask how large k should be in order that, with probability at least $1 - \epsilon$ we can correctly recover from these characters the topology of the underlying phylogenetic tree. Let $k_{\min}(\epsilon)$ be the smallest value of k that achieves this last property. Clearly $k_{\min}(\epsilon)$ depends on features of the generating tree, in particular the number n of leaves, and the assignment of $t(e)$ values to the edges of this tree (it also depends on ϵ, however we will regard this as a fixed small number). Any dramatic 'shortening' of an interior edge, or 'lengthening' of an exterior edge (i.e. making the $t(e)$ value small or large, respectively) will cause $k_{\min}(\epsilon)$ to diverge and so we will assume that each binary phylogenetic tree has all its $t(e)$ values in some fixed interval $[l_n, u_n]$ which may depend on n. The questions of interest are then to determine the dependence of $k_{\min}(\epsilon)$ on n and the values (l_n, u_n). Essentially this question provides another formalization of the question 'how much phylogenetic information is contained in characters that evolve according to a simple Markov model.' The authors of [13] showed that,

$$k_{\min}(\epsilon) \leq c' \cdot \frac{\log(n)}{l_n^2} \cdot \exp(u_n \delta(T)) \tag{14}$$

where c' is a constant (dependent only on ϵ) and $\delta(T)$ is a function (only) of the phylogenetic tree T that grows slowly with n. Specifically, $\delta(T)$ is at most a constant times $\log(n)$, but is typically (i.e. on average) $O(\log(\log(n))$.

Thus if we were to regard l_n and u_n as constants (independent of n) then $k_{\min}(\epsilon)$ is at worst polynomial in n, and more typically a power of $\log(n)$ (improving an alternative bound described in [16]). We have not mentioned the tree reconstruction method used to establish (14); it is a polynomial time (in $|X|$) algorithm, and chosen more for tractability of analysis than for any supposed superior performance; a comparable analysis for maximum likelihood seems more difficult [57].

An obvious question arises: is the bound on $k_{\min}(\epsilon)$ described by (14) and the consequent relationship between $k_{\min}(\epsilon)$ and n (for l_n, u_n fixed) optimal? Certainly $k_{\min}(\epsilon)$ must grow at least as fast as (a constant times) $\log(n)$, by elementary counting arguments. This applies under any

model of sequence evolution on a bounded state space and any tree reconstruction method [56]. The essence of this argument is the following: there are $\frac{(2n-4)!}{(n-2)!2^{n-2}}$ trivalent phylogenetic X–trees and r^{nk} sequences on an r–letter alphabet, and so if $k = o(\log(n))$ then for sufficiently large n there exist more trivalent phylogenetic X–trees than r–letter sequences of length k.

Also an inverse square dependence of $k_{\min}(\epsilon)$ on l_n is necessary, even when $n = 4$, as shown in [57]. However there is reason to believe that (14) is not optimal, provided that u_n is less than the critical transition value (viz. $\frac{1}{4}\log_e(2)$) between the ordered and unordered states, discussed above. This has led to the following conjecture, which promises a remarkable strengthening of (14) under a further restriction.

Conjecture 4.1. *Suppose that $u_n \leq u < \frac{1}{4}\log_e(2)$. Then*

$$k_{\min}(\epsilon) \leq c \cdot \frac{\log(n)}{l_n^2},$$

where c is a constant that depends only on ϵ and u.

Conjecture 4.1 is clearly true for trees for which $\delta(T)$ is bounded – these are trees for which no vertex is very far from a leaf. However for trees that have 'deep' vertices such as the complete balanced binary phylogenetic tree that has all its $n = 2^m$ leaves at distance m from a fixed central edge, bound (14) is polynomial in n. Yet precisely in this 'worst case' setting the bound promised by Conjecture 4.1 holds - this was recently established in [37], using an entirely different approach from [13]. The paper [37] also showed that the restriction on u_n is necessary for Conjecture 4.1 to hold, for when u_n is allowed to take larger values, polynomial dependence of $k_{\min}(\epsilon)$ on n can result.

Conjecture 4.1 has been extended to a much more general conjecture in [37] concerning the transition from logarithmic to polynomial dependence of $k_{\min}(\epsilon)$ on n for a range of Markov models at the corresponding transition from the ordered to unordered phase of the process. We give one formulation of this general conjecture below.

Suppose that $M(e) = \exp(t(e)Q)$ for some substitution rate matrix Q. Denote the eigenvalues of Q by $\{0 = \lambda_1, \lambda_2, \lambda_3 \ldots, \lambda_q\}$. Let $\eta = -\max\{\Re\lambda_2, \ldots \Re\lambda_q\}$, where $\Re(a + ib) = a$. Note that $\eta > 0$. Moreover, the absolute value of the second largest eigenvalue of $M(e)$ is given by $\exp(-t(e)\eta)$. We also know (by [42]) that if $2\exp(-t(e)\eta)^2 > 1$ for all e then the measure is in an ordered state. The condition above may be also written as $2\eta t(e) < \log_e(2)$ or $t(e) < \frac{\log_e(2)}{2\eta}$. We are thus led to the following general conjecture.

Conjecture 4.2. *Consider a model where for each edge e, $M(e) = \exp(t(e)Q)$. Suppose that $\ell_n \leq t(e) \leq u_n$ and that $u_n \leq u < \frac{\log_e(2)}{2\eta}$. Then*

$$k_{\min}(\epsilon) \leq c \cdot \frac{\log(n)}{l_n^2},$$

where c is a constant that depends only on ϵ and u.

Taking M to be the Jukes-Cantor model (2) and recalling (13), the conjecture asserts that if $\ell_n \leq t(e) \leq u_n < \frac{\log_e(2)}{6r}$ for all edges e, then a logarithmic number of characters suffices to

reconstruct the tree. Note that if $Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ is given by the CFN model, then $\eta = 2$ and therefore $\frac{\log_e(2)}{2\eta} = \frac{\log_e(2)}{4}$. So Conjecture 4.2 generalizes Conjecture 4.1.

4.2 Reconstructing forests

We have seen how a polynomial number of characters is needed when the underlying tree is 'deep' and substitution rates are such that the phylogenetic signal decays exponentially with time. However, in some cases, a more modest objective is posed - to find a 'large portion' of the underlying tree.

Note that for the rooted binary tree on 2^m leaves, where all the leaves are at distance m from the root, only an $O(2^{-r})$ fraction of the vertices is at distance r or more from the set of leaves. This is true in general for any binary trees - the proportion of vertices that are at separated by r or more edges from their nearest leaf is $O(2^{-r})$.

Since the 'deep' part of binary trees is small, the problem of reconstructing a 'large portion' of the tree may be much easier than the problem of reconstructing the complete tree. One way of reconstructing a 'large portion' of a tree is to reconstruct a forest containing many of the edges of the original tree.

When we reconstruct a phylogenetic forest, we partition the set of leaves X into α sets $X_1, \ldots, X_\alpha$ and for $1 \leq \beta \leq \alpha$ we reconstruct a tree T_β such that the leaf set of T_β is X_β. Moreover, the forest $(T_\beta)_{\beta \leq \alpha}$ has the following property. The forest may be obtained from T by removing $\alpha - 1$ edges; when we remove the edge (u, v) we contract the two edges adjacent to u to a single edge (provided u is not a leaf); similarly for v.

In [41] it is shown that a logarithmic number of characters suffices to reconstruct a forest containing most edges of the tree. The results in [41] are formulated in terms of the log-det distance. For simplicity, we formulate the results for the special case where Q is a symmetric substitution rate matrix corresponding to an ergodic process and $M(e) = \exp(t(e)Q)$ for all e.

Theorem 4.3. *Consider a binary phylogenetic tree T, where $M(e) = \exp(t(e)Q)$ and $\ell \leq t(e) \leq u$ for all e. Then for any $\epsilon > 0$ there exists a polynomial time reconstruction algorithm and a constant c (dependent only on ℓ, u and ϵ) such that for all $R > 0$ if $k \geq ce^{R+20u} \log n$ then given k characters generated independently by this process on T the algorithm outputs a partition of X into sets $X_1, \ldots, X_\alpha$ and constructs a forest $T_1, \ldots, T_\alpha$ such that, with probability at least $1 - \epsilon$, X_β is the set of leaves of T_β for all β and*

- *$T_1, \ldots, T_\alpha$ is a forest that may be obtained from T by removing $\alpha - 1$ edges;*
- *the number of trees α in the forest is at most $\lfloor 1 + \frac{60n}{\sqrt{2}} 2^{-\frac{R}{4u}} \rfloor$.*

Note in particular that taking $R = 4u(6 + r)$, the algorithm reconstructs a forest containing all but 2^{-r} of the edges of T. In particular, taking r to be a large constant, we may recover most edges of T from $O(\log n)$ characters.

The relationship between R and α shows as a function of the sequence length, parameterized by R, what level of refinement is achievable in reconstructing parts of the tree as parameterized by α.

5 Processes on an unbounded state space: The random cluster model

For the type of Markov model on a small state space that we have dealt with so far the subsets of the vertices of a phylogenetic tree T that are assigned particular states do not generally form connected subtrees of T (in biological terminology this is because of 'homoplasy' - the evolution of the same state more than once in the tree).

However increasingly there is interest in genomic characters such as gene order where the underlying state space may be very large ([19], [34], [35], [48]). For example, the order of k genes in a signed circular genome can take any of $2^k(k-1)!$ values. In these models whenever there is a change of state - for example a re-shuffling of genes by a random inversion (of a consecutive subsequence of genes) - it is likely that the resulting state (gene arrangement) is a unique evolutionary event, arising for the first time in the evolution of the genes under study. Indeed Markov models for genome rearrangement such as the (generalized) Nadeau-Taylor model [34], [45] confer a high probability that any given character generated is homoplasy-free on the underlying tree, provided the number of genes is sufficiently large relative to $|X|$ ([50]). In this setting a 'random cluster' model which we will describe here is the appropriate (limiting case) model, and may be viewed as the phylogenetic analogue of what is known in population genetics as the 'infinite alleles model' of Kimura and Crow [29].

Thus for this section we consider the size of the state space to be infinite (or at least very large, and perhaps variable with $|X|$). Some of the arguments described above are no longer valid in this setting. For example, the simple argument in Section 4.1 that showed that $k_{\min}(\epsilon)$ must grow at least as fast as the function $\log(n)$ does not apply when the size of state space is infinite, or finite but variable with $|X|$. Indeed it has recently been shown that for any trivalent phylogenetic X–tree T there is an associated set of just *four* characters for which T is the only phylogenetic X–tree on which each character in that collection has a homoplasy-free evolution (see [50], [26]). Thus it is reasonable to ask whether $O(1)$ characters might suffice to reconstruct T under a simple random model. We will see that the answer to this question is 'no', but clearly we need a different type of argument.

Consider the following random process on a phylogenetic tree T. For each edge e let us independently either cut this edge - with probability $p(e)$ - or leave it intact. The resulting disconnected graph (forest) G partitions the vertex set $V(T)$ of T into non-empty sets according to the equivalence relation that $u \sim v$ if v and v are in the same component of G. This model thus generates random partitions of $V(T)$, and thereby of X by connectivity, and we will denote these partitions of $V(T)$ and X using the symbols $\overline{\chi}$ and χ, respectively. Fig. 5(b) illustrates this process.

For an element $x \in X$ we will let $\chi(x)$ denote the equivalence class containing x. We call the resulting probability distribution on partitions of X the *random cluster model* with parameters (T, p) where p is the map $e \mapsto p(e)$.

In keeping with the biological setting we will call an arbitrary partition χ of X a *character* (on X). Let $\mathbb{P}[\chi|T,p]$ denote the probability of generating a character χ under the random cluster model with parameters (T,p). We say a subset C of the set $E(T)$ of edges of T is a *cutset for* χ *on*

Figure 3: (a) A trivalent phylogenetic X–tree T for $X = \{1, 2, \ldots, 7\}$; (b) For the random cluster model, cutting the edges of T that are marked by a cross induces the character χ on X given by $\chi = \{\{1,3\},\{2,4,5\},\{6\},\{7\}\}$.

T if the partition χ of X equals that induced by the components of $(V(T), E(T) - C)$. Then

$$\mathbb{P}[\chi|T,p] = \sum_{C} \prod_{e \in C} p(e) \prod_{e \in E(T)-C} (1-p(e)), \tag{15}$$

where the summation is over all cutsets C for χ on T. Note that the number of terms in the summation described by (15) can be exponential with $|X|$. However by modifying the well-known dynamic programming approach for computing the probability of a character on a tree according to a finite state Markov process (see eg. [17]) one can compute $\mathbb{P}[\chi|T,p]$ in polynomial time in $|X|$.

Suppose we generate a sequence $\Pi = (\chi_1, \ldots, \chi_k)$ of k such independent characters on X where the generating pair (T,p) is unknown. We wish to reconstruct T with probability at least $1-\epsilon$ from Π.

The following theorem, from [43] describes how the required value of k is related to the size of T and properties of p, and illustrates the logarithm-polynomial phase transition that occurs depending on whether or not the $p(e)$ values are all less than $1/2$. We refer the reader to [43] for the proof of this result.

Theorem 5.1. *Let $0 < l \leq u < 1$ and $0 < \epsilon < 1$ be fixed constants. Consider the random cluster model on any collection of the parameters (T,p) where T is a trivalent phylogenetic tree, and $l \leq p(e) \leq u$ for all edges e of T. Let k be the number of characters generated i.i.d. under this model, and $k_{\min}(\epsilon)$ be the minimal k such that the tree can be correctly reconstructed from the characters with probability at least $1-\epsilon$. Then, if n denotes the number of leaves of T.*

(i) $k_{\min}(\epsilon)$ grows logarithmically with n if $u < \frac{1}{2}$. In particular, if

$$k \geq \frac{2(1-u)^4}{l(1-2u)^4} \log\left(\frac{n}{\sqrt{\epsilon}}\right),$$

then the tree can be reconstructed correctly with probability $1-\epsilon$. Furthermore, there is a polynomial-time (in n) algorithm for reconstructing T from the generated characters.

(ii) $k_{\min}(\epsilon)$ *can grow polynomially with* n *if* $l > \frac{1}{2}$. *In particular, for all* h, *if*

$$k \leq \frac{\epsilon(1-l)^h}{6}\left(\frac{n}{3}\right)^{-\log_2(2-2l)}, \tag{16}$$

then there exists a distribution on trivalent phylogenetic X*-trees, such that if* T *is drawn according to the distribution,* $p(e) = l$, *for all edges of the trees, and characters are generated by* (T, p), *then the probability of correctly reconstructing* T *given the* k *characters is bounded above by* $\epsilon + 3^{-3\times 2^h}$.

Theorem 5.1 shows that the situation with the random cluster model differs from a bounded-state spaces model such as the CFN model, in two respects. Firstly, the critical value is $\frac{1}{2}$ instead of $\frac{1}{2}(1-\frac{1}{\sqrt{2}})$. This corresponds to the fact that in statistical physics models on the binary tree, the critical value for the extremality of the free measure or the Ising model is $\frac{1}{2}(1-\frac{1}{\sqrt{2}})$, see [5, 14, 39], while the critical value for uniqueness of Gibbs measure, or the critical value for percolation is $\frac{1}{2}$, see [23, 47]. In [40] it is shown that for any Markov model, if the substitution rate is high then k depends polynomially on n.

The second respect in which the random cluster model differs from the symmetric two state model, is that for the random cluster model, the dependence of k on l has exponent -1 rather than -2, see [13, 37, 57].

We now provide a brief outline of the proof of part (i) of Theorem 5.1, since it combines a combinatorial result with a probabilistic percolation argument; full details can be found in [43]. Recall that a *quartet tree* is a trivalent phylogenetic X–tree for $|X| = 4$. We can represent any quartet tree by the notation $xy|wz$ where x, y are leaves that are adjacent to one interior vertex, while w, z are leaves that are adjacent to the other interior vertex. For any trivalent phylogenetic X–tree, T let $\mathcal{Q}(T)$ denote the set of quartet trees induced by T by selecting subsets of X of size 4. It is a fundamental result from [9] that T is uniquely determined by $\mathcal{Q}(T)$. Suppose that T is a trivalent phylogenetic X–tree. We say that T *displays* a quartet tree $xy|wz$ (respectively, a set $\mathcal{Q}$ of quartet trees) if $xy|wz \in \mathcal{Q}(T)$ (respectively, if $\mathcal{Q} \subseteq \mathcal{Q}(T)$). For example the tree T in Fig. 5(a) displays the quartet tree 12|47. For any three distinct vertices a, b, c of T let $\text{med}(a, b, c)$ denote the *median vertex* of the triple a, b, c; that is, the unique vertex of T that is shared by the paths connecting a and b, a and c and b and c.

A collection $\mathcal{Q}$ of quartet trees is a *generous cover* of T if $\mathcal{Q} \subseteq \mathcal{Q}(T)$ and if, for all pairs of interior vertices u, v there exists a quartet tree $xx'|yy' \in \mathcal{Q}$ for which $u = \text{med}(x, x', v)$ and $v = \text{med}(u, y, y')$. Given a sequence $\mathcal{C} = (\chi_1, \chi_2, \ldots, \chi_k)$ of characters on X, let

$$\mathcal{Q}(\mathcal{C}) = \{xx'|yy' : \exists i \in \{1, \ldots, k\} : \chi_i(x) = \chi_i(x') \neq \chi_i(y) = \chi_i(y')\}.$$

One of the main steps in the proof of Theorem 5.1 is to establish the following purely combinatorial result:

Proposition 5.2. *If* $\mathcal{Q}$ *is a generous cover of a trivalent phylogenetic* X*–tree* T *then* T *is the only phylogenetic* X*–tree that displays* $\mathcal{Q}$.

Using a percolation-style argument, one can then show that provided k is at least as large as that specified in Theorem 5.1(i), an i.i.d sequence $\mathcal{C}$ of k characters will, with probability at least $1 - \epsilon$ have the property that $\mathcal{Q}(\mathcal{C})$ is a generous cover of T.

A further extension of Proposition 5.2, along with a simulation-based study of the random cluster model has been described recently by [11].

6 Large but finite state spaces

Finally, we turn to the question of how many characters one needs to reconstruct a large tree if the characters evolve under a Markov model on a large but finite state space. The results of this section are based on [55] where further details can be found.

As mentioned earlier, many processes involving simple reversible models of change can be modelled by a random walk on a regular graph. To explain this connection, suppose there are certain 'elementary moves' that can transform each state into some 'neighboring' states. In this way we can construct a graph from the state space, by placing an edge between state α and state β precisely if it is possible to go from either state to the other in one elementary move. The graph so obtained is said to be *regular*, or more specifically *d–regular* if each state is adjacent to the same number d of neighboring states.

For example, aligned sequences of length N under the r–state Poisson model can be regarded as a random walk on the set of all sequences of length N over R; here an elementary move involves changing the state at any one position to some other state (chosen uniformly at random from the remaining $r-1$ states). Thus the associated graph has r^N vertices and it is N–regular.

As another example, consider a simple model of (unsigned) genome rearrangement where the state space consists of all permutations of length N (corresponding to the order of genes $1, \ldots, N$) and an elementary move consists of an inversion of the order of the elements of the permutation between positions i and j, where this pair is chosen uniformly at random from all such pairs between $\{1, \ldots, N\}$. In this case the state space has size $N!$ and the graph is d–regular for $d = \binom{N}{2}$.

Both of the graphs we have just described have more structure than mere d–regularity. To describe this we recall the concept of a Cayley graph. Suppose we have a (non-abelian or abelian) group $\mathcal{G}$ together with a subset S of elements, with the properties that $1_{\mathcal{G}} \notin S$ and $s \in S \Rightarrow s^{-1} \in S$. Then the *Cayley graph* associated with the pair $(\mathcal{G}, S)$ has vertex set $\mathcal{G}$ and an edge connecting g and g' whenever there exists some element $s \in S$ for which $g = g' \cdot s$. To recover the above graph on aligned sequences of length N over an r–letter alphabet, we may take $\mathcal{G}$ as the (abelian) group $(\mathbb{Z}_r)^N$ and the set S of all N–tuples that are the identity element of $\mathbb{Z}_r$ except on one co-ordinate. To recover the graph described above for unsigned genome rearrangements we may take $\mathcal{G}$ to be the (nonabelian) symmetric group on N letters and S to be the elements corresponding to inversions.

The demonstration that such graphs are Cayley graphs has an important consequence - it implies that they also have the following property. A graph G is said to be *vertex-transitive* if, for any two vertices u and v there is an automorphism of G that maps u to v. Informally, a graph is vertex-transitive if it 'looks the same, regardless of which vertex one is standing at'. Clearly a (finite) vertex-transitive graph must be d–regular for some d, and it is an easy and standard exercise to show that every Cayley graph is vertex-transitive (however not every vertex-transitive graph is a Cayley graph, and not every regular graph is vertex-transitive).

Suppose that R is a group, and for some subset S (closed under inverses and not containing the identity element of R) we have $Q_{\alpha\beta} = q$ if and only if there exists some element $s \in S$ for which

$\beta = \alpha \cdot s$, otherwise for any distinct pair α, β we have $Q_{\alpha\beta} = 0$. Such a process we will call a *group walk process (on the generating set S).* Group walk processes have a useful property on trees: for each arc $e = (u, v)$ of $T = (V, E)$ consider the event $\Delta(e)$ that the state that occurs at v is different from the state that occurs at u (i.e. there has been a net transition across the edge). Then using the fact that the Cayley graph for (R, S) is vertex transitive, it is easily shown that the events $(\Delta(e), e \in E)$ are independent.

For these models we then ask how likely it is that a character evolves without homoplasy on a tree. This question has been investigated for the 2–state Poisson model (and pairs of taxa) by Chang and Kim [6]. Here we consider more general processes on a larger state space, and for many taxa - consequently we obtain bounds rather than the exact expressions that are possible in the simpler setting of [6]. The proof of the following Lemma is straightforward (for details, see [55]).

Lemma 6.1. *Let $(X_t; t \geq 0)$ be a group walk process on generating set of size d. Then, for any two distinct states α, β, and any values $s, t \geq 0$,*

$$\mathbb{P}[X_{t+s} = \beta | X_t = \alpha] \leq \frac{1}{d}.$$

The following result shows that for a group-walk process if the size of the generating set is much larger than $2n^2$ (where n is the number of species) then any character generated on a tree with n species will almost certainly be homoplasy-free on that tree.

Proposition 6.2. *Suppose characters evolve on a phylogenetic tree T according to a group walk process on a generating set of size d. Let $p(T)$ denote the probability that the resulting randomly-generated character χ is homoplasy-free on T. Then*

$$p(T) \geq 1 - \frac{(2n-3)(n-1)}{d}$$

where $n = |X|$.

Proof. Consider a general Markov process on T with state space R. Suppose that for each arc (u, v) of T and each pair α, β of distinct states in R, the conditional probability that state β occurs at v given that α occurs at u is at most p. Then, from [50] (Proposition 7.1) we have $p(T) \geq 1 - (2n-3)(n-1)p$. By Lemma 6.1 we may take $p = \frac{1}{d}$. The result now follows. □

We are now ready to state a result for certain Markov processes on large (but finite!) state spaces, which brings together several ideas presented above. Informally, Theorem 6.3 states that for a group walk process, a growth of around $n^2 \log(n)$ in the size of the generating set is sufficient (with all else held constant) for producing a sequence of homoplasy-free characters that define T.

Let $p_{\min} = \min\{\mathbb{P}[\Delta(e)] : e \in E\}$, $p_{\max} = \max\{\mathbb{P}[\Delta(e)] : e \in E\}$, and for any $\epsilon > 0$ let

$$c_\epsilon = \frac{1 + \log(\frac{1}{\sqrt{\epsilon}})}{\beta\epsilon} \tag{17}$$

where $\beta = p_{\min}(\frac{1-2p_{\max}}{1-p_{\max}})^4$.

Theorem 6.3. *Suppose characters evolve i.i.d. on a binary phylogenetic tree T according to a group walk process on a generating set of size d, where*

$$d \geq c_\epsilon \cdot n^2 \log(n)$$

with c_ϵ given by (17). Then with probability at least $1-2\epsilon$ we can correctly reconstruct the topology of T by generating an appropriate (and logarithmic in n) number of characters.

Proof. Let us generate $k = \lceil \frac{2}{\beta} \log(\frac{n}{\sqrt{\epsilon}}) \rceil$ characters under a group walk process on a rooted phylogenetic tree. Consider the event E that all of these characters are homoplasy-free on T. By Proposition 6.2 we have

$$\mathbb{P}[E] \geq (1 - \frac{(2n-3)(n-1)}{d})^k \geq (1 - \frac{2}{c_\epsilon \log(n)})^{\frac{2}{\beta}\log(\frac{n}{\sqrt{\epsilon}})}.$$

Now, applying the inequality $(1-x)^y \geq 1 - xy$ for $x, y > 0$, and straightforward algebra gives

$$\mathbb{P}[E] \geq 1 - \frac{\epsilon(1+\log(\frac{1}{\sqrt{\epsilon}}))^{-1}}{\log(n)}[\log(n) + \log(\frac{1}{\sqrt{\epsilon}})] \geq 1 - \epsilon,$$

using $\log(n) \geq 1$.

To relate the group walk process to the random cluster model we use a simple coupling argument. First consider any linear extension of the partial order induced by T on its vertices - i.e. impose a time-scale on the tree. To each vertex v assign the pair (α, i) where α is the state of the group walk process at v, and $i \in \{0, 1, 2, ...\}$ indicates how often this state has arisen from another state earlier in the tree. Note that this model always generates new states (it is homoplasy-free); furthermore, transition events on different edges of the tree are independent. Consequently this coupled process is precisely the random cluster model, and we may identify β with the expression $a(\frac{1-2b}{1-a})^4$ in Theorem 5.1(i). Furthermore the probability that T will be correctly reconstructed from k characters produced by the coupled process is at least $1-\epsilon$ by Theorem 5.1.

On the other hand the original k characters induce the same partitions as the derived characters whenever event E holds, and $\mathbb{P}[E] \geq 1 - \epsilon$. Consequently, by the Bonferroni inequality, the joint probability that event E holds and that the k characters produced by the coupled process recovers T is at least $1 - 2\epsilon$. Thus the probability that the original k characters recover T is at least this joint probability, and so at least $1 - 2\epsilon$, as claimed. □

References

[1] Aldous, D. and Fill, J. A. (2003). *Reversible Markov chains and random walks on graphs*, book in preparation. Current version available at http://stat-www.berkeley.edu/users/aldous/book.html.

[2] Alon, N. and Spencer, J. H. (2000). *The probabilistic method*, second edition. John Wiley and Sons.

[3] Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*, Springer-Verlag.

[4] Bandelt, H. J. and Dress, A. W. M. (1986). Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, **7**, 309–343.

[5] Bleher, P. M., Ruiz, J. and Zagrebnov, V. A. (1995). On the Purity of limiting Gibbs state for the Ising model on the Bethe lattice. *J. Stat. Phys.*, **79**, 473–482.

[6] Chang J. T. and Kim, J. (1996). The measurement of homoplasy: a stochastic view. In *Homoplasy: The recurrence of similarity in evolution*, (ed. M.J. Sanderson and L. Hufford), 189–303.

[7] Charleston, M. and Steel, M. A. (1995). Five surprising properties of parsimoniously colored trees. *Bulletin of Mathematical Biology*, **57(2)**, 367–375.

[8] Churchill, G. A., von Haesler, A. and Navidi, W. C. (1992). Sample size for a phylogenetic inference. *Mol. Biol. Evol.*, **9**, 753–769.

[9] Colonius, H., and Schulze, H. H. (1981). Tree structures for proximity data. *British Journal of Mathematical and Statistical Psychology*, **34**, 167–180.

[10] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley and Sons.

[11] Dezulian, T. and Steel, M. (2004). Phylogenetic closure operations and homoplasy-free evolution. *Proceedings of the International Federation of Classification Societies (IFCS)*. In press.

[12] Diaconis, P. (1988). Group representations in probability and statistics. *Institute of Mathematical Statistics, Lecture Notes-Monograph Series*, **Vol 11**, (ed. Gupta, S. S.). Hayward, California.

[13] Erdös, P. L., Székely, L. A., Steel, M. and Warnow, T. (1999). A few logs suffice to build (almost) all trees (I). *Random Structures and Algorithms*, **14**, 153–184.

[14] Evans, W., Kenyon, C., Peres, Y. and Schulman, L. J. (2000). Broadcasting on trees and the Ising Model. *Annals of Applied Probability*, **10(2)**, 410–433.

[15] Evans, S. N. and Speed, T. P. (1993). Invariants of some probability models used in phylogenetic inference. *Annals of Statistics*, **21**, 355–377.

[16] Farach, M. and Kannan, S. (1999). Efficient algorithms for inverting evolution. In *Journal of the Association for Computing Machinery*, **46**, 437–449.

[17] Felsenstein, J. (1981a). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

[18] Felsenstein, J. (2004). *Inferring Phylogenies*, Sinauer Press.

[19] Gallut, C. and Barriel, V. (2002). Cladistic coding of genomic maps. *Cladistics*, **18**, 526–536.

[20] Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, **18**, 866–873.

[21] Georgii, H. O. (1988). Gibbs measures and phase transitions. *de Gruyter Studies in Mathematics*, Walter de Gruyter and Co., Berlin, 9.

[22] Georgii, H. O., Häggstrm, O. and Maes, C. (2001). The random geometry of equilibrium phases. *Phase Transitions and Critical Phenomena*, (eds. Domb, C. and Lebowitz, J. L.), Academic Press, London. pp. 1–142.

[23] Grimmett, G. (1999). *Percolation* (2nd edn). Springer-Verlag, Berlin.

[24] Guiasu, S. (1977). *Information theory with applications*, McGraw-Hill, New York.

[25] Higuchi, Y. (1977). Remarks on the limiting Gibbs states on a $(d+1)$-tree. *Publications of the Research Institute for Mathematical Sciences*, **13(2)**, 335–348.

[26] Huber, K. T., Moulton, V. and Steel, M. (2002). Four characters suffice to convexly define a phylogenetic tree. *Research Report UCDMA2002/12*, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand.

[27] Janson, S. and Mossel, E. (2003). Robust reconstruction on trees is determined by the second eigenvalue. to appear in *Annals of Probability*

[28] Kesten, H. and Stigum, B. P. (1966). Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Annals of Mathematical Statistics*, **37**, 1463–1481.

[29] Kimura, M. and Crow, J. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.

[30] Lecointre, G., Philippe, H., Lanh van Le, H. and Le Guyader, H. (1994). How many nucleotides are required to resolve a phylogenetic problem? *Molecular phylogenetics and evolution*, **3**, 292–309.

[31] Martin, J. (2003). Reconstruction thresholds on regular trees. In *Discrete Random Walks*, (eds. Banderier, C. and Krattenthaler, C.), *Discrete Mathematics and Theoretical Computer Science*, 191–204. Availible at `http://dmtcs.loria.fr/proceedings/dmACind.html`.

[32] Martinelli, F., Sinclair, A. and Weitz, D. (2003). The ising model on trees: Boundary conditions and mixing time. Submitted to *Communication in Mathematical Physics*. Extended abstract appeared in [33].

[33] Martinelli, F., Sinclair, A. and Weitz, D. (2003). The ising model on trees: Boundary conditions and mixing time. In *Proceedings of the Forty Fourth Annual Symposium on Foundations of Computer Science*, 628–639.

[34] Moret, B. M. E., Tang, J., Wang, L.S. and Warnow, T. (2002). Steps toward accurate reconstruction of phylogenies from gene-order data. *Journal of Computer and System Sciences*, **65(3)**, 508–525.

[35] Moret, B. M. E., Wang, L. S., Warnow, T. and Wyman, S. (2001). New approaches for reconstructing phylogenies based on gene order. Proc. 9th Int'l Conf. on Intelligent Systems for Molecular Biology ISMB-2001, *Bioinformatics*, **17**, S165–S173.

[36] Mossel, E. (2001). Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, **11(1)**, 285–300.

[37] Mossel, E. Phase transitions in phylogeny. *Transactions of the American Mathematical Society*, in press.

[38] Mossel, E. (2001). Reconstruction on trees: Beating the second eigenvalue. *Annals of Applied Probability*, **11**, 285–300.

[39] Mossel, E. (1998). Recursive reconstruction on periodic trees. *Random Structures and algorithms*, **13**, 81–97.

[40] Mossel, E. (2003). On the impossibility of reconstructing ancestral data and phylogenies. *Journal of computational biology*, **10(5)**, 669–676.

[41] Mossel, E. (2004). Distorted metrics on trees and phylogenetic forests. submitted.

[42] Mossel, E and Peres, Y. (2003). Information flow on trees. *Annals of Applied Probability*, **13(3)**, 817–844.

[43] Mossel, E. and Steel, M. (2003). A phase transition for a random cluster model on phylogenetic trees. *Mathematical Biosciences*, (in press).

[44] Pemantale, R. and Peres, Y. (1995). Recursions on trees and the ising model at critical temperatures. Unpublished manuscript.

[45] Nadeau, J. J. and Taylor, B. A. Lengths of chromosome segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences, USA*, **81**, 814–818.

[46] Penny, D., McComish, B.J., Charleston, M. A. and Hendy, M. D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution*, **53**, 711–723.

[47] Peres, Y. (1997). Probability on trees: an introductory climb. *Lectures on probability theory and statistics (Saint-Flour, 1997)*, 193–280. Lecture Notes in Math., 1717, Springer, Berlin.

[48] Rokas, A. and Holland, P. W. H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution*, **15**, 454–459.

[49] Schröder, E. (1870). Vier combinatorische probleme. *Zeitschrift für Mathematik und Physik*, **15**, 361–376.

[50] Semple, C. and Steel, M. (2002). Tree reconstruction from multi-state characters. *Advances in Applied Mathematics*, **28**, 169–184.

[51] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press.

[52] Sober, E. and Steel, E. (2002). Testing the hypothesis of common ancestry. *Journal of Theoretical Biology*, **218**, 395–408.

[53] Steel, M. (1994). Recovering a tree from the leaf colourations it generates under Markov model. *Applied Mathematical Letters*, **7**, 19–23.

[54] Steel, M. (2001). My favourite conjecture, http://www.math.canterbury.ac.nz/~mathmas/conjecture.pd

[55] Steel, M. and Penny, D. (2004). MP and the phylogenetic information in multi-state characters. (submitted).

[56] Steel, M. and Székely, L. A. (1999). Inverting random functions (I). *Annals of Combinatorics*, **3**, 103–113.

[57] Steel, M. and Székely, L. A. (2002). Inverting random functions (II): explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM Journal on Discrete Mathematics*, **15**, 562–575.

[58] Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. (1996). Phylogenetic inference. In *Molecular Systematics* (2nd edn.), (eds. Hillis, D. M, Moritz, C. and Marble, B.K.), Sinauer, Sunderland, U.S.A., 407–514.

[59] Tuffley, C. and Steel, M. (1998). Modelling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, **147**, 63–91.